Review

Clinical Text Data in Machine Learning: Systematic Review

Irena Spasic[1], PhD; Goran Nenadic[2], PhD

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
[2]Department of Computer Science, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Irena Spasic, PhD
School of Computer Science and Informatics
Cardiff University
5 The Parade
Cardiff, CF24 3AA
United Kingdom
Phone: 44 02920870320
Email: spasici@cardiff.ac.uk

Abstract

Background: Clinical narratives represent the main form of communication within health care, providing a personalized account of patient history and assessments, and offering rich information for clinical decision making. Natural language processing (NLP) has repeatedly demonstrated its feasibility to unlock evidence buried in clinical narratives. Machine learning can facilitate rapid development of NLP tools by leveraging large amounts of text data.

Objective: The main aim of this study was to provide systematic evidence on the properties of text data used to train machine learning approaches to clinical NLP. We also investigated the types of NLP tasks that have been supported by machine learning and how they can be applied in clinical practice.

Methods: Our methodology was based on the guidelines for performing systematic reviews. In August 2018, we used PubMed, a multifaceted interface, to perform a literature search against MEDLINE. We identified 110 relevant studies and extracted information about text data used to support machine learning, NLP tasks supported, and their clinical applications. The data properties considered included their size, provenance, collection methods, annotation, and any relevant statistics.

Results: The majority of datasets used to train machine learning models included only hundreds or thousands of documents. Only 10 studies used tens of thousands of documents, with a handful of studies utilizing more. Relatively small datasets were utilized for training even when much larger datasets were available. The main reason for such poor data utilization is the annotation bottleneck faced by supervised machine learning algorithms. Active learning was explored to iteratively sample a subset of data for manual annotation as a strategy for minimizing the annotation effort while maximizing the predictive performance of the model. Supervised learning was successfully used where clinical codes integrated with free-text notes into electronic health records were utilized as class labels. Similarly, distant supervision was used to utilize an existing knowledge base to automatically annotate raw text. Where manual annotation was unavoidable, crowdsourcing was explored, but it remains unsuitable because of the sensitive nature of data considered. Besides the small volume, training data were typically sourced from a small number of institutions, thus offering no hard evidence about the transferability of machine learning models. The majority of studies focused on text classification. Most commonly, the classification results were used to support phenotyping, prognosis, care improvement, resource management, and surveillance.

Conclusions: We identified the data annotation bottleneck as one of the key obstacles to machine learning approaches in clinical NLP. Active learning and distant supervision were explored as a way of saving the annotation efforts. Future research in this field would benefit from alternatives such as data augmentation and transfer learning, or unsupervised learning, which do not require data annotation.

(JMIR Med Inform 2020;8(3):e17984) doi: 10.2196/17984

KEYWORDS

natural language processing; machine learning; medical informatics; medical informatics applications

Introduction

Clinical narratives represent the main form of communication within health care. In comparison with generically coded elements of electronic health records (EHRs), the narrative notes provide a more detailed and personalized account of patient history and assessments, offering a better context for clinical decision making [1]. Natural language processing (NLP) is a subfield of artificial intelligence that studies the ways in which the analysis and synthesis of information expressed in a natural language can be automated. It has repeatedly demonstrated its feasibility to unlock evidence buried in clinical narratives, making it available for large-scale analysis down the stream [2]. Traditionally, rule-based approaches were commonly used to unlock evidence of specific types [3]. Their development requires some form of direct interaction with clinical experts to convert their knowledge, often tacit, into a set of explicit pattern-matching rules.

Machine learning has long been hailed as a silver bullet solution for the knowledge elicitation bottleneck, the main argument being that the task of annotating the data manually is easier than that of eliciting the knowledge [4]. Nonetheless, the amount of data required to train a machine learning model may require as much time to annotate as the knowledge elicitation itself [5]. Much like the law of energy conservation, it seems that the knowledge required to inform the creation of an accurate computational model is simply transferred from one form to another. Instead of explicit knowledge in the form of rules, machine learning is based on implicit knowledge in the form of annotations and their distribution, with the time involved in their acquisition remaining virtually constant.

Another problem associated with the machine learning approach is the availability of clinical narratives given the sensitive nature of health data and privacy concerns [6]. These problems (ie, unavailability of manually annotated data) may result in the lack of representativeness of the training data and consequently substandard performance of the corresponding machine learning models. For these reasons, the main aim of this review was to provide systematic evidence on the properties of data used to train machine learning approaches to clinical NLP. In addition, we investigate the types of NLP tasks that have been supported by machine learning and how they can be applied in clinical practice.

The remainder of the paper is organized as follows. We start by explaining the methodology of this systematic review in detail. We then discuss the main findings of the review. Finally, we conclude the review by outlining future research directions in this field.

Methods

Overview

On the basis of the guidelines for performing systematic reviews described by Kitchenham [7], our methodology is structured around the following steps. First, research questions (RQs) were used to define the scope, depth, and the overall aim of the review. Next, a search strategy was designed to identify all studies that are relevant to the RQs in an efficient and reproducible manner. In addition, inclusion and exclusion criteria were defined to refine the scope. A critical appraisal of the included studies was conducted to ensure that the findings of the review are valid. During data extraction, the relevant information was identified from the included studies and semistructured to facilitate the synthesis of evidence and support the findings of the review.

Research Questions

The overarching topic of this review is concerned with the properties of text data used to enable machine learning approaches to clinical NLP. The main aim of the review was to answer the RQs given in Table 1. RQ1 aims at describing the properties of data that are relevant for interpreting the performance of machine learning. These properties include size, provenance, heterogeneity, annotations, and others. Here, heterogeneity refers to content, structure, and clinical domains. RQ2 classifies the problems addressed by machine learning in the context of NLP into different types of computational tasks. Finally, RQ3 focuses on the ways in which NLP based on machine learning can be applied to tackle practical problems encountered in clinical practice.

Table 1. Research questions.

ID	RQ[a]
RQ1	What are the key properties of data used to train and evaluate machine learning models?
RQ2	What types of NLP[b] tasks have been supported by machine learning?
RQ3	How can NLP based on machine learning be applied in clinical practice?

[a]RQ: research question.

[b]NLP: natural language processing.

Search Strategy

We used PubMed as a search engine to retrieve relevant documents from the MEDLINE database of 28 million citations from life sciences and biomedical literature, which are indexed by Medical Subject Headings (MeSH). MeSH is a hierarchically organized controlled vocabulary used for manually indexing articles in MEDLINE in a uniform and consistent manner to facilitate their retrieval. We derived a list of search terms to describe the topic of this review: *machine learning, deep learning, text, natural language, clinical, health, health care,* and *patient*. Here, machine learning and deep learning are used

to retrieve articles that employ this methodology. Note that MeSH includes the term *machine learning*, thus making it unnecessary to include specific machine learning techniques such as *support vector machines* or *conditional random fields* into the search query. The following 2 search terms, *text* and *natural language*, refer to the relevant type of input into the learning methods. The final 4 terms were used to refer to clinical applications. Owing to the broad nature and common use of the last 6 terms, their mentions were restricted to titles and abstracts only. In an attempt to prevent retrieval of nonoriginal studies and NLP applications developed to support systematic reviews, we negated the terms *literature*, *bibliometric*, and *systematic review*. Finally, to focus on the emerging application of machine learning, we restricted the search to the period from January 1, 2015. The search was performed on August 8, 2018. The search terms were combined into a PubMed query as follows:

((“machine learning”[All Fields] OR “deep learning”[All Fields]) AND (text[Title/Abstract] OR “natural language” [Title/Abstract]) AND (clinical[Title/Abstract] OR health [Title/Abstract] OR healthcare[Title/Abstract] OR patient [Title/Abstract]) NOT (literature[Title/Abstract] OR bibliometric [Title/Abstract] OR “systematic review”[Title/Abstract]) AND (“2015/01/01”[PDat] : “2018/08/08”[PDat])

We identified 389 candidate articles according to the described search strategy. The results were further screened against the selection criteria.

Selection Criteria

The scope of this systematic review was formally defined by the inclusion and exclusion criteria given in Textboxes 1 and 2, respectively. Having screened the retrieved articles against the inclusion and exclusion criteria, a total of 149 articles were retained for further processing.

Textbox 1. Inclusion criteria.

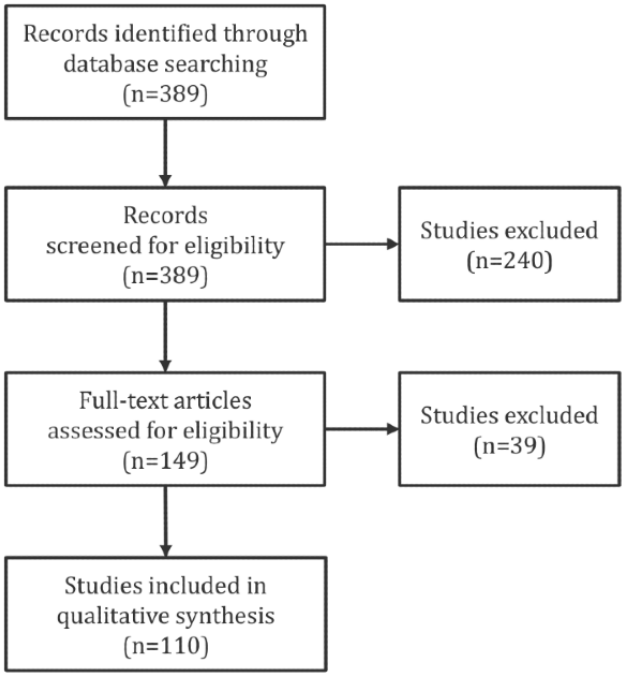
1. The study has to use natural language processing.
2. Machine learning has to be used to support such processing.
3. Input text has to be routinely collected within health care boundaries.
4. Input text has to be written or dictated.
5. The article has to be peer reviewed.
6. The full text has to be freely available online for academic use.

Textbox 2. Exclusion criteria.

1. Articles written in a language other than English.
2. Natural language processing of a language other than English.
3. Natural language processing of spoken language.

Given the interdisciplinary nature of articles considered for this review, we encountered a wide diversity of venues in which they were published. Not surprisingly, some studies put an emphasis on the clinical aspects but neglected to describe the computational aspects of the study in sufficient detail to support its reproducibility. To be included in this review, articles needed to provide sufficient information to support answering RQs defined in Table 1. In other words, they needed to describe the datasets used; define the NLP problem clearly; describe the features used to support NLP; state the machine learning methods used and, where appropriate, their parameters; and provide a formal evaluation of the results. A total of 39 studies were found not to match these criteria. This further reduced the number of selected articles to 110 [8-117]. Figure 1 summarizes the outcomes of the 4 major stages in the systematic literature review.

Figure 1. Flow diagram of the literature review process.



Data Extraction

We explored the selected studies to extract data that contribute to answering the RQs given in Table 1. Data were extracted from the full text of articles under the following headings: data, task, clinical domain, and clinical application. The data properties considered included their size, provenance, collection methods, annotation, and any relevant statistics. The task was defined as a subfield of NLP (eg, text classification, information extraction (IE), named entity recognition (NER), and word sense disambiguation [WSD]). This was supplemented with task-specific information; for example, for NER, we also specified the type of named entities considered. Clinically relevant information was extracted to identify the potential for practical applications. The extracted data were then used to facilitate a narrative synthesis of the main findings.

Results

The first step in developing a machine learning model is to collect data relevant to the problem at hand. Ultimately, the model's performance will depend on the properties of such a dataset. We summarized these properties, including data size, key data sources, training annotations, and types of clinical documents considered.

Size

Among other factors, the performance of machine learning models and the significance of test results depend on the size of the dataset used for training and testing, respectively. In this section, we examine the size of datasets used in the studies included in this review. Owing to large variations in data sizes, we used a logarithmic scale to fit this information into the chart shown in Figure 2, which stratifies the datasets according to their order of magnitude. Some studies used as few as 40 documents [48] and as few as 15 patients [28]. The vast majority of datasets have the cardinality in the range of hundreds or thousands. Only 10 studies used tens of thousands of documents, with a handful of studies utilizing more than that despite the fact that machine learning approaches are data hungry in the sense that their performance is strongly correlated with the amount of training data available.

Relatively small datasets were utilized even when much larger datasets were available. Figure 3 demonstrates data utilization on a logarithmic scale, with some studies utilizing as little as 0.002% of available data [44] and as much as 11.88% [11]. Specific examples illustrate this issue: 500 from 188,843 [32], 300 from 4025 [59], 62 from 6343 [25], 323 from 16,000 [24], 1188 from 10,000 [11], 1610 from 52,746 [39], 1004 from 96,303 [112], 1058 from 376,487 [34], 10,000 from 103,564 sentences [36], less than 12,000 out of 137,522+28,159 [101], 562 from 2.5 million [44], 8288 from 2,977,739 [13], 6174 from 2.6 million [113], 3467 from 8,168,330 [68], and 2159 from 24 million [19].

Figure 2. Distribution of data size on a logarithmic scale.

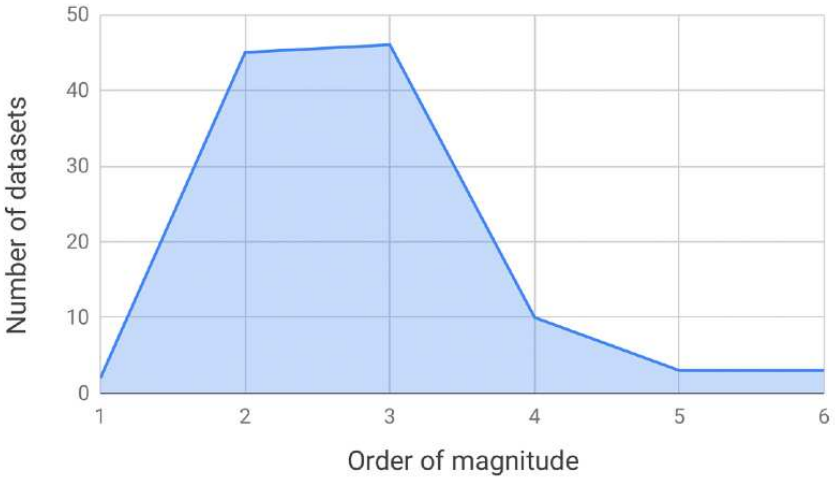
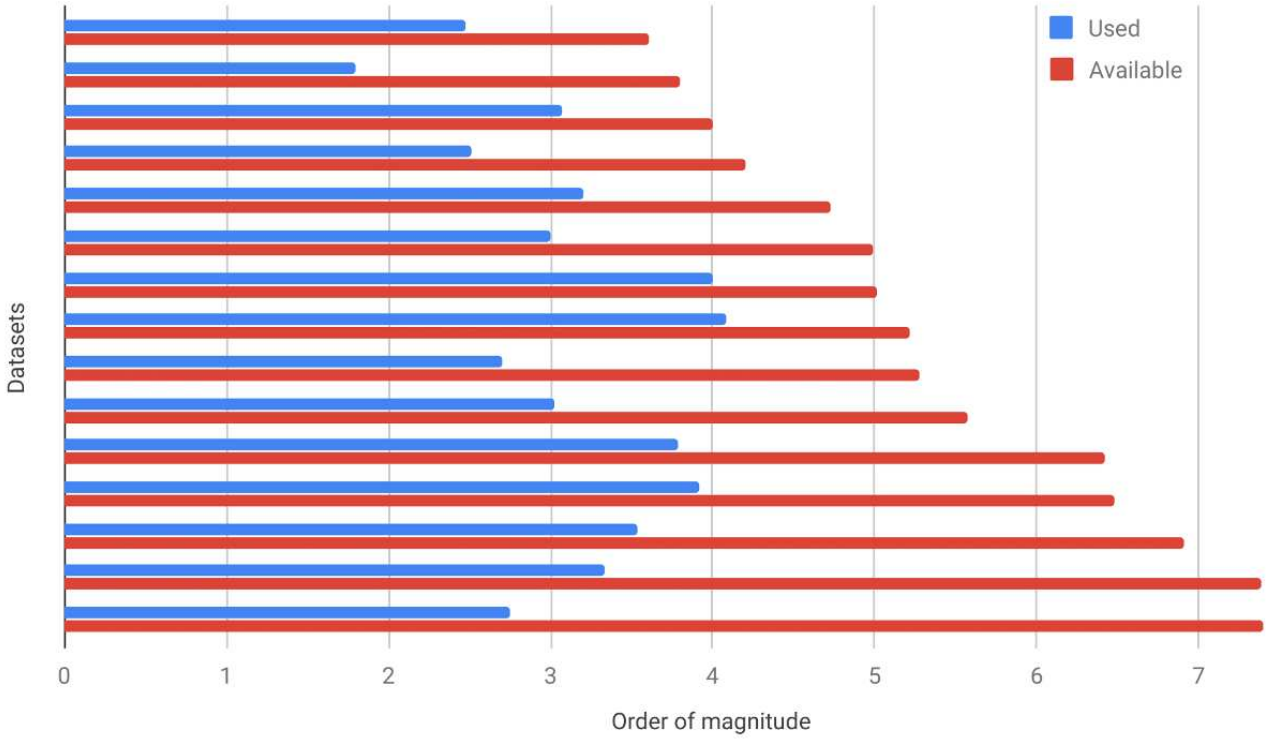


Figure 3. Data utilization on a logarithmic scale.



Annotation

The main reason for such poor data utilization is the annotation bottleneck faced by supervised machine learning algorithms, which require training data to be annotated to generalize them into predictive mathematical models. Compiling manually annotated corpora is both labor-intensive and error prone. The fact that annotations are task-specific means that the training data rarely get to be recycled. The labor and time limitations imposed on individual studies will naturally be correlated with the volume of manually annotated training data. Active learning aims to address the annotation bottleneck by involving human experts in the machine learning process in an attempt to improve performance with relatively small annotation effort [20,54,100]. An active learning algorithm can iteratively sample a subset of data for manual annotation, depending on the current predictive performance. Sampling strategies can be based on a disagreement between different predictive models or different measures of uncertainty, density, and expectation of a single predictive model. Such sampling depends on the quality of a predictive model and may not be efficient when retraining the model lasts relatively long. Alternatively, diversity measures can be used to prioritize annotation. For instance, pair-wise cosine similarity was used to compare sentences and prioritize those least similar to annotated sentences for annotation [20]. However, this may lead to the selection of outliers, whose presence in the training data can result in a degradation of predictive models. By considering representativeness and informativeness, outliers are less likely to be selected, thus leading to better coverage of the data characteristics and, consequently better predictive models. Here, the average similarity between a sentence and all other sentences indicates

how representative it is [54]. The higher the similarity, the more representative the sentence is.

In principle, supervised learning approaches are convenient when labels are readily available. For instance, EHRs combine different types of data elements from unstructured data such as free text and images to structured data (ie, discrete elements such as numbers, dates, and codes) from controlled medical terminologies [118]. In the studies included in this systematic review, larger datasets (ie, those ranging from tens of thousands to millions, see Figure 2), were used mostly in cases where existing structured data were utilized as labels. For instance, in relation to hospitalization, readily available information about events such as in-hospital death [102], discharge [90], readmission [9], and emergency department visits [37] was used to train models to predict future events of this type well in advance to inform an appropriate course of action. Similarly, in relation to diagnostics, both prior (eg, imaging protocol [17,94]) and posterior (eg, test result [69]) information was utilized for supervision. International Classification of Diseases (ICD) diagnosis codes were used to train predictive models from historical data to identify patients at risk [16,22,50] or to facilitate disease surveillance [76]. Similarly, supervised models trained with ICD procedure codes otherwise used for billing can be used for cost optimization but also improving the quality of care [81]. Indeed, all of these examples have clear applications in care improvement and resource management. In some other cases (eg, classification of clinical notes into medical subdomains [103]), the utility of such information remains unclear.

Some types of learning problems such as WSD lend themselves well to semiautomated labeling based on greedy matching. Not surprisingly, the corresponding methods were tested on large datasets [33,105]. Similarly, using the concept of distant supervision, which utilizes an existing knowledge base to automatically annotate raw text, as much as 9.5 million clinical notes were annotated with adverse drug events [99]. Where manual annotation was unavoidable, crowdsourcing was explored. This approach is suitable for patient-facing problems such as readability of medical documents [116], where lay annotators are indeed ideally suited for the annotation task. The concept of crowdsourcing was explored for problems that require medical expertise [24]. Even though the interannotator agreement among crowdsourced workers was found to be much lower than that of medical experts, with Krippendorff alpha coefficient over .7, it was still found to be good agreement beyond chance. However, privacy constraints do undermine the feasibility of crowdsourcing in the context of clinical narratives.

Provenance

Besides the small volume of training data, another issue that might affect the performance of machine learning methods trained on such data is their provenance. The structure and style of clinical narratives may vary greatly between institutions [119]. Therefore, when the provenance of data is confined to a small number of contributing institutions, the data may not be representative. This, in turn, may lead to overfitting, a modeling error that occurs when a complex model adapts to the idiosyncrasies of the training data and fails to generalize the underlying properties of the problem. Unfortunately, the majority of studies reviewed here were limited to the authors' host institutions [8,10,12,15,17,22,24,25,28,30-33,35,40,41,44,66,70,76,79,84-86,89,90,94,95,99,105,106,111,113]. Rarely are such datasets freely accessible to the community. A notable exception is the Medical Information Mart for Intensive Care (MIMIC) [120], a freely accessible critical care database that stores a wide range of clinical narratives, including radiology reports [87], clinical notes [102] and discharge summaries [16,39]. Although it is a single-site dataset, some consolation may be found in the sheer volume of data. More importantly, its public availability allows for rigorous and detailed direct comparison of competing approaches, a rare commodity in clinical NLP.

Only 9 studies used data from 2 institutions [36,47,50,56,61,100,103,109,112]. Three studies used data from 3 institutions [45,71,87]. A handful of studies managed to obtain data from multiple sources: 5 [38], 6 [73], 18 [19], and 28 [37]. The Veterans Health Administration (VHA) [121,122], as the largest integrated health care system in the United States, provides centralized access to data from multiple institutions, enhancing the credibility of results achieved on such data [13,14,29,34,55,68,72,77,97].

Availability

Most datasets used in the included studies originated from a few institutions, thus offering no hard evidence about the transferability of machine learning models. Knowing that the format and style of clinical notes may vary substantially across institutions [119], it is not uncommon to observe a significant drop in performance when training a model in one institution and testing it in another [33,61,75,105,109]. This remains an ongoing concern for the clinical NLP community, where the confidentiality of data involved requires a careful balance between accessibility and privacy protection. In this section, we discuss wider availability of data that provide opportunities for secondary uses, including research. In this context, the NLP community challenges play an important role in providing access to clinical data to a wider pool of researchers and establishing benchmarks for future comparisons. Not surprisingly, many studies reviewed here have been enabled by the datasets shared in community challenges, which are described in Table 2.

Table 2. Datasets used in clinical natural language processing community challenges.

Dataset	Provenance	Documents	Size[a]	Annotations	Studies
2010 i2b2/VA [123]	PHC[b], BIDMC[c], UPMC[d]	Discharge summaries, progress reports	871	Medical problems, treatments, tests, and relations among them	[20,49,64,67,96,104]
2011 i2b2/VA [124]	PHC, BIDMC, UPMC, Mayo[e]	Discharge summaries, progress reports, radiology reports, pathology reports, other reports	978+164	Coreference chains for the problem, person, test, result, treatment, anatomical site, disease or syndrome, sign or symptom, etc	[63]
2012 i2b2 [125]	PHC, BIDMC	Discharge summaries	310	Clinical events, temporal expressions, temporal relations	[64]
2013 ShARe/CLEF eHealth [126]	BIDMC	Discharge summaries, electrocardiogram reports, echocardiogram reports, radiology reports	300	Disorders, acronyms, and abbreviations	[54,57,88,98,114]
2014 i2b2/UTHealth [127,128]	PHC	Longitudinal medical records	1304	Protected health information; risk factors for heart disease	[18,21,26,52,62,64,80,82,91,107,108]
2015 SemEval/THYME [129]	Mayo	Clinical notes, pathology reports	600	Times, events, and temporal relations among them	[60]
2016 CEGS N-GRID [130,131]	PHC	Psychiatric intake records	1000	Protected health information; symptom severity	[23,27,42,53,58,65,78,83,92]

[a]Size is expressed as the number of documents.

[b]Partners Health Care (PHC) is a nonprofit hospital and physician network that includes Brigham and Women's Hospital and Massachusetts General Hospital.

[c]Beth Israel Deaconess Medical Center (BIDMC) is a teaching hospital of Harvard Medical School. Both organizations are based in Boston, Massachusetts, United States.

[d]The University of Pittsburgh Medical Center (UPMC) is a global nonprofit health enterprise that integrates over 35 hospitals, 600 clinical locations, and a health insurance division.

[e]The Mayo Clinic is a nonprofit academic medical center based in Rochester, Minnesota, which focuses on integrated clinical practice, education, and research. The clinic specializes in treating difficult cases through tertiary care.

Similarly, MIMIC dataset represents a key driver of open research in clinical NLP. It is notable for being the only freely accessible critical care database of its kind [120]. Data analysis is unrestricted once a data use agreement is accepted, enabling clinical research and education internationally. The open nature of the data supports the reproducibility of findings and enables continual research advances. MIMIC is a large, single-center database that stores deidentified, comprehensive clinical information relating to patients admitted to critical care units at the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, United States, a large tertiary care hospital. Its content, which spans more than a decade, integrates different types of data (see Table 3). Of interest to this systematic review are free-text data, which include various types of notes and reports. Their integration with coded data offers an opportunity to circumvent manual annotation of data for supervised learning and evaluation purposes. For instance, Berndorfer and Henriksson [16] used a large dataset of 59,531 discharge summaries with at least one assigned ICD diagnosis code to automate the process of diagnosis coding. However, in many cases, accurate classification of medical conditions exists only in clinical narratives. Therefore, it may be necessary to annotate relevant phrases in the free text to train classification models. For instance, Gehrmann et al [39] manually annotated 1610 discharge summaries from MIMIC to automatically learn which phrases are relevant for 10 patient phenotypes considered. Similarly, Tahmasebi et al [87] manually annotated 860 radiology reports from MIMIC and 2 other institutions to evaluate an unsupervised approach to detecting and normalizing anatomical phrases.

Table 3. Description of clinical data types in the Medical Information Mart for Intensive Care.

Type	Description
Billing	Coded data recorded primarily for billing and administrative purposes.
Descriptive information	Demographic information, admission and discharge times, and dates of death.
Dictionaries	Look-up tables for cross-referencing identifiers (eg, codes) with associated definitions.
Interventions	Procedures such as dialysis, imaging studies, and placement of lines.
Laboratory measurements	Blood chemistry, hematology, urine analysis, and microbiology test results.
Medications	Administration records of intravenous medications and medication orders.
Notes	Free-text notes such as provider progress notes and hospital discharge summaries.
Physiologic information	Nurse-verified vital signs, approximately hourly (eg, heart rate, blood pressure, and respiratory rate).
Reports	Free-text reports of electrocardiogram and imaging studies (x-ray, computed tomography, ultrasound, and magnetic resonance imaging).

In addition to openness, an important driver of advancing state of the art in clinical NLP is an ability to access a wide range of data sources, many of which may not be compatible with national or organization-wide standards. As the largest integrated health care system in the United States, which provides care at 1243 health care facilities, including 172 medical centers and 1062 outpatient sites of care of varying complexity, the VHA [121,122] has the potential to address this challenge. The VHA offers veterans (ie, those who served in the active military, naval, or air service and who were discharged or released under conditions other than dishonorable) a wide range of inpatient, outpatient, mental health, rehabilitation, and long-term care services, which are all linked by an EHR platform. The construction of the VHA's information infrastructure, the Veterans Information Systems Technology Architecture (VistA), began in 1982 and became operational in 1985. VistA integrates multiple applications seamlessly that are accessible via a graphical user interface, the Computerized Patient Record System, first launched in 1997. Designed primarily to support clinical care delivery rather than billing, the system has been used since 2004 to document all routine clinical activities currently storing more than 16 billion clinical entries.

On average, 1 million free-text notes (eg, progress notes and discharge summaries), 1.2 million provider-entered electronic orders, 2.8 million images (radiologic studies, electrocardiograms, and photographs), and 1 million vital signs were stored in VistA daily. Such proliferation of data quickly outgrew the original plans for storage capacity, network bandwidth, support staff, and information technology budget, leading to the construction of the Corporate Data Warehouse (CDW) in 2006. The new repository for patient-level data aggregated from across the VHA's national health delivery system also hosts data from the legacy system, each featuring its own data rules, definitions, and structures. Given the slow process of normalizing these idiosyncrasies to a common standard and the rapidly increasing volume of data, the CDW allowed selective streaming of data from VistA and structuring them pragmatically in a way that minimizes redundancy. The CDW stores comprehensive patient-level data, which are used primarily to support health care delivery, but their unprecedented richness and volume provide a great opportunity for secondary uses such as quality improvement and research. To facilitate such uses, the VHA has partitioned a section of the CDW for use by health services and informatics investigators, who can access these data in secure workspaces within the VHA's firewall. The VHA is developing mechanisms to fully deidentify data extracts so that they can be shared outside of the VHA.

Similar to MIMIC, integration of structured (coded) and unstructured (free-text) data offers an opportunity to circumvent manual annotation of data for supervised learning and evaluation purposes. In this manner, Ben-Ari et al [14] utilized postoperative notes of 32,636 patients by cross-referencing them to prescription data. However, most studies still rely on manual annotation of information that is not well documented in structured data. For example, Bates et al [13] manually annotated 8288 radiology reports as *fall* or *not fall* at the document level. Similarly, Maguen et al [68] annotated 3467 randomly selected psychotherapy notes with respect to the use of evidence-based psychotherapy. Patterson et al [77] manually annotated 2000 colonoscopy procedure notes with an indication, which included screening, nonscreening, noncolonoscopy, and unknown. Walsh et al [97] annotated 3900 snippets of text referring to axial spondyloarthritis in a corpus sampled from 500 million clinical notes and 120 million radiology notes. Divita et al [29] sampled 948 records from 164 preselected document types and annotated them manually to identify 5819 positively asserted symptoms within the documents. Kim et al [55] annotated a corpus of 1465 echocardiography reports, radiology reports, and other note types from multiple medical centers sampled at random for mentions and assessments of left ventricular ejection fraction. Fodeh et al [34] sampled 1058 clinical notes of 101 types and manually annotated fine-grained information about pain assessments, which included not only pain mention but also its features such as intensity, quality, site, and etiology. Meystre et al [72] sampled a cohort of 1083 patients and annotated their clinical notes of more than 10 preselected types with information regarding congestive heart failure treatment performance measures. These in-document annotations were summarized at the clinical note and patient level for binary classification of patients as meeting the treatment performance measure or not. These studies illustrate the extent of manual annotation effort involved in developing machine learning approaches to clinical NLP. Unfortunately, manual annotations remain underexploited

because the fruits of such labor are rarely shared outside the original teams of investigators.

Types of Narratives

The vast majority of studies focused on a single type of clinical narrative. This may be driven by a specific clinical application. For instance, Mai and Krauthammer [69] focused exclusively on free-text test orders to predict whether a patient would test positive for a particular virus in a quest to reduce viral testing volumes. To support service improvement, Elmessiry et al [30] focused solely on patient complaints. Similarly, applications related to patient safety focused on relevant documents such as adverse event reports [15], patient safety event reports [35], and incident reports [101].

Not surprisingly, most clinical applications of NLP focus on diagnosis and prognosis as they are central to medicine. Clinicians and health policymakers need to make predictions about the diagnosis and disease prognosis to support their decision making. These 2 applications focus primarily on various types of reports. For instance, electroencephalography reports were used to study epilepsy [41,70], whereas echocardiography reports were used to extract information relevant to cardiovascular medicine [55]. Most studies explored radiology reports [13,24,43,45,85,87,110,111]. They typically focus on a single imaging modality such as computer tomography [11,48,71,106,112] or magnetic resonance imaging (MRI) [17,47,94]. Such a segregated approach may be warranted by the intrinsic differences in the types of images produced, which may be reflected in the types of information discussed in the corresponding reports. For instance, MRI better differentiates between soft tissues than x-ray imaging does. Therefore, their respective reports may focus on different types of anatomical structures and their pathologies. This implies that machine learning models trained on one type of report may not be transferrable to another.

Nonetheless, aggregating findings from multiple imaging modalities [19,46,73] and other types of examination may increase diagnostic accuracy, especially when planning surgical treatments. In particular, pathology and radiology form the core of cancer diagnosis, leading to an initiative to integrate pathology and radiology studies to support making correct diagnoses and appropriate patient management and treatment decisions [132]. In this context, Bahl et al [10] combined mammographic reports, image-guided core needle biopsy reports, and surgical pathologic reports to avoid unnecessary surgical excisions. An important data source that supports this type of integration is RadBank**,** a database that links radiology and pathology reports [133]. It contains more than 2 million reports and allows full- text search by patient history, findings, and diagnosis by radiology and pathology. Still, the majority of studies focused on pathology reports alone [8,22,38,66,75,76]. Combinations of different report types were mostly used in enabling studies that focused on NLP tasks without a specific clinical application in mind (eg, NER approaches trained on electrocardiography, echocardiography, and radiology reports) [54,57,88,98,114].

Heterogeneity across different types of reports, including cardiac catheterization procedure reports, coronary angiographic reports together with integrated reports that combine history and physical report, discharge summary, outpatient clinic notes, outpatient clinic letter, and inpatient discharge medication report retrieved from the Emory Cardiovascular Biobank [134] was utilized to train robust machine learning models [115]. Different subsets drawn from clinical notes, admission notes, discharge summaries, progress reports, radiology reports, allergy entries, and free-text medication orders are typically used to support fundamental NLP applications such as spell-checking [56]; coreference resolution [63]; WSD [100], including that of abbreviations [105]; and NER [20,64]. Finally, colonoscopy reports were used to explore the feasibility of NLP in a clinical setting [77,93].

Discharge summaries are used as the primary communication means between hospitals and primary care and, as such, are essential for ensuring patient safety and continuity of care. Their content and structure may vary greatly between institutions and clinicians [135]. Typical components include dates of admission and discharge, reason for hospitalization, significant findings from history and examination, significant laboratory findings, significant radiological findings, significant findings from other tests, list of procedures performed, procedure report findings, stress test report findings, pathology report findings, discharge diagnosis, condition at discharge, discharge medications, follow up issues, pending test results, and information provided to patients and family. Practically, discharge summaries may be viewed as amalgamations of different types of clinical narratives, some of which we discussed previously. Although this may make their processing more challenging, any algorithms trained on discharge summaries are more likely to be applicable across a wider range of clinical narratives. Discharge summaries tend to provide the most informative accounts of patient phenotypes and have been used to automate cohort selection [39]. This also makes them well suited for training and testing NER approaches [59,96,104], extraction of relationships between them [49,67], or predicting diagnoses [16].

Other types of clinical narratives considered include physician notes [84], progress notes [25,40,90], EHR notes [74,81,116], surgical notes [14,79], and emergency department notes [50,109]. Unspecified type of clinical notes [102] were used mostly for classification [9,12,31,61,86,95,103,113], WSD [33], and disambiguation and IE [36,51,99].

Psychiatric notes were used mainly in an NLP community challenge to extract protected health information and symptom severity [23,27,42,53,58,65,78,83,92]. These narratives are key enablers of mental health informatics as the fine-grained context of actionable information does not readily lend itself to predefined coding schemes. Other types of documents used to support mental health applications include psychotherapy notes [68], event and correspondence notes [32], progress notes [40], and those in general clinical context including admission notes and discharge summaries [117].

Longitudinal EHRs were mainly used in NLP community challenges [18,21,26,52,62,80,82,91,107,108]. In practical applications, cumulative patient profiles were used to predict frequent emergency department visits [37]. Longitudinal records consisting of encounter and clinical notes were used to determine

whether a candidate problem is genuine or not [28]. Similarly, encounter notes were used to determine whether a specific dermatological problem was definite, probable, or negative [44].

Clinical Applications

This section focuses on the clinical applications of NLP approaches based on machine learning. We mapped 21 clinical applications against 7 NLP tasks (see Figure 4). It should be noted that we excluded a total of 39 studies that did not provide sufficient information to support answering RQs defined in Table 1. These studies may have described their own clinical applications, which are not discussed in this section.

Figure 4. Clinical applications underpinned by natural language processing tasks.

	Classification	Clustering	Coreference resolution	Information extraction	Named entity recognition	Ranking	Word sense disambiguation	Total
Care improvement	8			2				10
Comparative effectiveness	1							1
Data management	2			1	1			4
Diagnosis	2							2
Efficiency	1							1
Enabling	7		2	4	14		3	30
Interactive NLP	1			1				2
Knowledge acquisition				1				1
Patient literacy						1		1
Pharmacovigilance	3			1				4
Phenotyping	13							13
Prognosis	13			3				16
Quality	2							2
Referral	1							1
Resource management	8							8
Risk prediction	1							1
Safety	4							4
Service improvement	1							1
Surveillance	5			1	1			7
Triage	1	2		1	1			5
Unclear	1							1
Total	**75**	**2**	**2**	**15**	**17**	**1**	**3**	

Not surprisingly, the vast majority of studies focused on the task of text classification, which naturally lends itself to supervised machine learning. Most commonly, the classification results were used to support phenotyping, prognosis, care improvement, resource management, and surveillance.

EHR-based phenotyping approaches leverage data collected routinely in the course of health care delivery to identify cohorts of individuals that share certain clinical characteristics, events, and service patterns [136]. Their data can then be used for the secondary purposes of observational and interventional studies, prospective recruitment into clinical trials, health services research, public health surveillance, and comparative effectiveness research. Standardized computable phenotypes can enable large-scale studies while ensuring reliability and reproducibility. For instance, historical trial patient enrollment decisions were used to demonstrate the potential of NLP to increase trial screening efficiency by 450% and reduce workload associated with patient cohort identification by 90% [137]. Different types of events identified from EHRs include falls [13] and long bone fractures [43]. Most often, EHR phenotyping focused on a single medical condition, eg, axial spondyloarthritis [97], hypertension [89], systemic lupus erythematosus [95], dermatitis [44], obesity [61], celiac disease [22], epilepsy [41], autism [84], or psychiatric problems in general [40]. Two studies differentiated between multiple disorders. Tran and Kavuluru [92] focused on 11 mental disorders including attention-deficit hyperactivity disorder, anxiety, bipolar disorder, dementia, depression, eating disorder, grief, obsessive compulsive spectrum disorder, psychosis, and posttraumatic stress disorder. Gehrmann et al [39] focused on a less homogeneous list of 10 disorders including advanced cancer, advanced heart disease, advanced lung disease, chronic neurologic dystrophies, chronic pain, alcohol abuse, substance abuse, obesity, psychiatric disorders, and depression.

In terms of prognosis, text classification results were used to predict 3-month survival [12], the likelihood of intracranial hemorrhage [11] and the development of coronary artery disease [18,26,52,62,80,82,91,107,108] or prognosis based on cancer staging [75].

At the other end of the spectrum from text classification were lower-level tasks such as coreference resolution [63,110] and WSD [33,100,105], which were not associated with any particular clinical application. However, their importance lies in enabling other higher-level NLP tasks. Similarly, as a subtask of IE, NER can be used to support structuring text into predefined templates, whose slots need to be filled with named entities of relevant types. The majority of NER studies were related to NLP community challenges such as those described in studies by Uzuner et al [123], Suominen et al [126], and Stubbs et al [131]. They focused on entities such as medical problems, tests, and treatments [20,49,67,96,104]; disorders [54,57,88,98,114]; and protected health information [27,58,65].

Unlike NER, the more complex task of IE found a wider variety of clinical applications, the most prominent of which include prognosis and care improvement. For instance, cancer stage is one of the most important prognostic parameters in cancer, but this information is typically recorded in clinical narratives, which means that medical abstractors have to read through large volumes of text to extract such information. Given the importance and laboriousness of this task, it is not a coincidence that all IE approaches with prognosis as the most obvious clinical application focused on cancer staging [8,38,111]. Another IE approach related to cancer focused on extraction of symptoms experienced by patients during chemotherapy [36]. Rather than prognosis, this information can be used to improve patient care through modifying treatments and recognizing and managing symptoms. Similarly, extraction of information about assessments and medications can be used to improve management and outpatient treatment of patients suffering from chronic heart failure [72].

Triage is a process for sorting patients into groups based on their need for or likely benefit from medical treatment. Clustering, which is the task of grouping objects in a way that objects within a cluster are more similar to one another than to those in other clusters, can, therefore, naturally be applied to triage patients. Clustering was used to identify latent groups of lymphoma patients from their pathology reports [66]. Another study confirmed that automatically generated clusters of radiology reports coincided with major topics in radiology investigations [46]. Surprisingly, triage was not found to be a common clinical application of NLP and was largely associated with a single author [45-48].

Summary

In this review, we examined the key properties of data used to train and evaluate machine learning models. We found that the size of the training dataset tends to be relatively small. For instance, the vast majority of studies included only hundreds or thousands of documents. Relatively small proportions were utilized for training even when much larger datasets were available. Beside their volume being small, training data were typically sourced from few institutions. In addition to the NLP community challenges such as i2b2, ShARe/CLEF eHealth, and CEGS N-GRID, most commonly used data sources were MIMIC and VHA. The vast majority of studies focused on a single type of clinical narratives, which ranged from imaging reports to hospital discharge summaries. Most often, training data were used to support the tasks of text classification, IE, and NER. Only a handful of studies focused on tasks such as clustering, ranking, coreference resolution, and WSD. Most commonly, the classification results were used to support clinical applications such as phenotyping, prognosis, care improvement, resource management, and surveillance. The remaining NLP tasks did not have clear clinical applications. In fact, the majority were used to enable other higher-level NLP tasks.

Discussion

The use of text data in health informatics applications present quite a few challenges, the main ones being the preservation of patient privacy and the annotation bottleneck. Consequently, the training datasets become inflicted with problems typically associated with an unrepresentative sample. In other words, they may not reflect the distribution of characteristics of the target problem. In machine learning, such bias may lead to overfitting, a modeling error that occurs when a complex model adapts to idiosyncrasies of the training data and fails to generalize the underlying properties of the problem.

Unfortunately, most datasets used in the included studies originated from few institutions, thus offering no hard evidence about the generalizability and transferability of machine learning models. With the format and style of clinical notes varying substantially across institutions [119], a significant drop in performance was observed when training a model in one institution and testing it in another [33,61,75,105,109]. In this context, NLP community challenges play an important role in providing access to clinical data to a wider pool of researchers and establishing benchmarks for future comparisons. Not surprisingly, many studies included in this systematic review were enabled by the datasets shared in NLP community challenges. Unfortunately, relying on these challenges to provide clinical text data to NLP researchers seems like putting a Band-Aid on a proverbial bullet wound. Alternative opportunities have presented themselves in the form of synthetic health data, which contain the health records of realistic albeit not real patients. For instance, Synthea, the original open source synthetic health data software, can be used to simulate disease progression and the corresponding medical care to produce risk-free health care records at scale [138]. As synthetic data are not associated with any privacy concerns, crowdsourcing remains an option for their annotation though it may still require medical expertise, which remains an expensive commodity.

In terms of data annotation, lessons can be learned from other fields such as computer vision and speech processing, which have similarly been plagued by the lack of annotated data. They use data augmentation techniques to diversify data available for training of machine learning models without actually collecting any new data [139]. Similar techniques are now increasingly used to augment text data in a quest to improve generalization performance of the corresponding machine learning models [140-143]. Alternatively, transfer learning can be applied to leverage knowledge (features, parameters, etc) acquired in one domain and/or task with sufficient training data to support learning in another, which has got significantly less training data, thereby reducing expensive data annotation efforts [144,145]. In some cases, manual data annotation can be avoided altogether by applying the concept of distant supervision, which relies on an existing knowledge base to annotate text data automatically [146].

Some problems (eg, in-hospital death [102], discharge [90], readmission [9], and emergency department visits [37]), where labels are readily available, lend themselves naturally to supervised learning approaches. For instances, EHRs combine free-text data with codes from controlled medical terminologies, which can be utilized as class labels [118]. These codes were used to train predictive models from historical data to identify patients at risk [16,22,50], facilitate disease surveillance [76], or optimize the cost and quality of care [81]. For other problems, where data have to be annotated manually from scratch, insisting

on supervised learning is very much like trying to fit a square peg through a round hole, leaving unsupervised approaches such as topic modeling largely underexplored even though they may be better fit for purpose for clinical applications such as EHR phenotyping, patient triage, care, and service improvement.

In summary, we identified the data annotation bottleneck as one of the key obstacles to machine learning approaches in clinical NLP. Active learning has been explored as a way of using the annotation efforts in a more strategic manner. However, the clinical NLP community could benefit from using alternatives such as data augmentation, transfer learning, and distant supervision. Ultimately, unsupervised learning avoids the need for data annotation altogether and, therefore, should be used more frequently to support clinical NLP.

Acknowledgments

The authors gratefully acknowledge the support from the Engineering and Physical Sciences Research Council for HealTex—UK Healthcare Text Analytics Research Network (Grant number EP/N027280/1).

Conflicts of Interest

None declared.

References

1. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
2. Spasi I, Uzuner O, Zhou L. Emerging clinical applications of text analytics. Int J Med Inform 2020 Feb;134:103974. [doi: 10.1016/j.ijmedinf.2019.103974] [Medline: 31630961]
3. Spasi I, Livsey J, Keane JA, Nenadi G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-623 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.009] [Medline: 25008281]
4. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition (Morgan Kaufmann Series in Data Management Systems). Burlington, Massachusetts, USA: Morgan Kaufmann; 2008.
5. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009 Mar;24(2):8-12. [doi: 10.1109/mis.2009.36]
6. Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med 2002;26(1-2):25-36. [doi: 10.1016/s0933-3657(02)00050-7] [Medline: 12234715]
7. Kitchenham B. Department of Computer Science: Keele University. 2004. Procedures for Performing Systematic Reviews URL: http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf [accessed 2020-03-24]
8. AAlAbdulsalam A, Garvin J, Redd A, Carter M, Sweeny C, Meystre S. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. AMIA Jt Summits Transl Sci Proc 2018;2017:16-25 [FREE Full text] [Medline: 29888032]
9. Agarwal A, Baechle C, Behara R, Zhu X. A natural language processing framework for assessing hospital readmissions for patients with COPD. IEEE J Biomed Health Inform 2018 Mar;22(2):588-596. [doi: 10.1109/JBHI.2017.2684121] [Medline: 28328520]
10. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology 2018 Mar;286(3):810-818. [doi: 10.1148/radiol.2017170549] [Medline: 29039725]
11. Banerjee I, Madhavan S, Goldman R, Rubin D. Intelligent word embeddings of free-text radiology reports. AMIA Annu Symp Proc 2017;2017:411-420 [FREE Full text] [Medline: 29854105]
12. Banerjee I, Gensheimer MF, Wood DJ, Henry S, Aggarwal S, Chang DT, et al. Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives. Sci Rep 2018 Jul 3;8(1):10037 [FREE Full text] [doi: 10.1038/s41598-018-27946-5] [Medline: 29968730]
13. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. J Am Med Inform Assoc 2016 Apr;23(e1):e113-e117 [FREE Full text] [doi: 10.1093/jamia/ocv155] [Medline: 26567329]
14. Ben-Ari A, Chansky H, Rozet I. Preoperative opioid use is associated with early revision after total knee arthroplasty: a study of male patients treated in the veterans affairs system. J Bone Joint Surg Am 2017 Jan 4;99(1):1-9. [doi: 10.2106/JBJS.16.00167] [Medline: 28060227]
15. Benin AL, Fodeh SJ, Lee K, Koss M, Miller P, Brandt C. Electronic approaches to making sense of the text in the adverse event reporting system. J Healthc Risk Manag 2016 Aug;36(2):10-20. [doi: 10.1002/jhrm.21237] [Medline: 27547874]
16. Berndorfer S, Henriksson A. Automated diagnosis coding with combined text representations. Stud Health Technol Inform 2017;235:201-205. [doi: 10.3233/978-1-61499-753-5-201] [Medline: 28423783]
17. Brown A, Marotta T. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. J Am Med Inform Assoc 2018 May 1;25(5):568-571. [doi: 10.1093/jamia/ocx125] [Medline: 29092082]

18. Buchan K, Filannino M, Uzuner O. Automatic prediction of coronary artery disease from clinical narratives. J Biomed Inform 2017 Aug;72:23-32 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.019] [Medline: 28663072]
19. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 2017 May;69:177-187 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.011] [Medline: 28428140]
20. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform 2015 Dec;58:11-18 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.010] [Medline: 26385377]
21. Chen Q, Li H, Tang B, Wang X, Liu X, Liu Z, et al. An automatic system to identify heart disease risk factors in clinical texts over time. J Biomed Inform 2015 Dec;58(Suppl):S158-S163 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.002] [Medline: 26362344]
22. Chen W, Huang Y, Boyle B, Lin S. The utility of including pathology reports in improving the computational identification of patients. J Pathol Inform 2016;7:46 [FREE Full text] [doi: 10.4103/2153-3539.194838] [Medline: 27994938]
23. Clark C, Wellner B, Davis R, Aberdeen J, Hirschman L. Automatic classification of RDoC positive valence severity with a neural network. J Biomed Inform 2017 Nov;75S:S120-S128 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.005] [Medline: 28694118]
24. Cocos A, Qian T, Callison-Burch C, Masino AJ. Crowd control: effectively utilizing unscreened crowd workers for biomedical data annotation. J Biomed Inform 2017 May;69:86-92 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.003] [Medline: 28389234]
25. Cohen KB, Glass B, Greiner HM, Holland-Bouley K, Standridge S, Arya R, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed Inform Insights 2016;8:11-18 [FREE Full text] [doi: 10.4137/BII.S38308] [Medline: 27257386]
26. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform 2015 Dec;58(Suppl):S53-S59 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.029] [Medline: 26210359]
27. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Learning to identify protected health information by integrating knowledge- and data-driven algorithms: a case study on psychiatric evaluation notes. J Biomed Inform 2017 Nov;75S:S28-S33 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.005] [Medline: 28602908]
28. Devarakonda MV, Mehta N, Tsou C, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform 2017 Sep;105:121-129 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.015] [Medline: 28750905]
29. Divita G, Luo G, Tran LT, Workman TE, Gundlapalli AV, Samore MH. General symptom extraction from VA electronic medical notes. Stud Health Technol Inform 2017;245:356-360. [Medline: 29295115]
30. Elmessiry A, Cooper WO, Catron TF, Karrass J, Zhang Z, Singh MP. Triaging patient complaints: Monte Carlo cross-validation of six machine learning classifiers. JMIR Med Inform 2017 Jul 31;5(3):e19 [FREE Full text] [doi: 10.2196/medinform.7140] [Medline: 28760726]
31. Fan Y, Zhang R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):51 [FREE Full text] [doi: 10.1186/s12911-018-0626-6] [Medline: 30066648]
32. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep 2018 May 9;8(1):7426 [FREE Full text] [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]
33. Finley G, Pakhomov S, McEwan R, Melton G. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. AMIA Annu Symp Proc 2016;2016:560-569 [FREE Full text] [Medline: 28269852]
34. Fodeh SJ, Finch D, Bouayad L, Luther SL, Ling H, Kerns RD, et al. Classifying clinical notes with pain assessment using machine learning. Med Biol Eng Comput 2018 Jul;56(7):1285-1292 [FREE Full text] [doi: 10.1007/s11517-017-1772-1] [Medline: 29280092]
35. Fong A, Harriott N, Walters DM, Foley H, Morrissey R, Ratwani RR. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. Int J Med Inform 2017 Aug;104:120-125. [doi: 10.1016/j.ijmedinf.2017.05.005] [Medline: 28529113]
36. Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage 2018 Jun;55(6):1492-1499. [doi: 10.1016/j.jpainsymman.2018.02.016] [Medline: 29496537]
37. Frost DW, Vembu S, Wang J, Tu K, Morris Q, Abrams HB. Using the electronic medical record to identify patients at high risk for frequent emergency department visits and high system costs. Am J Med 2017 May;130(5):601.e17-601.e22. [doi: 10.1016/j.amjmed.2016.12.008] [Medline: 28065773]
38. Gao S, Young M, Qiu J, Yoon H, Christian J, Fearn P, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2018 Mar 1;25(3):321-330. [doi: 10.1093/jamia/ocx131] [Medline: 29155996]

39. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 2018;13(2):e0192360 [FREE Full text] [doi: 10.1371/journal.pone.0192360] [Medline: 29447188]
40. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health 2017 Aug;20(3):83-87 [FREE Full text] [doi: 10.1136/eb-2017-102688] [Medline: 28739578]
41. Goodwin T, Harabagiu S. Multi-modal patient cohort identification from EEG report and signal data. AMIA Annu Symp Proc 2016;2016:1794-1803 [FREE Full text] [Medline: 28269938]
42. Goodwin TR, Maldonado R, Harabagiu SM. Automatic recognition of symptom severity from psychiatric evaluation records. J Biomed Inform 2017 Nov;75S:S71-S84 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.020] [Medline: 28576748]
43. Grundmeier RW, Masino A, Casper T, Dean J, Bell J, Enriquez R, Pediatric Emergency Care Applied Research Network. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2016 Nov 9;7(4):1051-1068 [FREE Full text] [doi: 10.4338/ACI-2016-08-RA-0129] [Medline: 27826610]
44. Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. IEEE Int Conf Healthc Inform 2017 Aug;2017:83-90 [FREE Full text] [doi: 10.1109/ICHI.2017.31] [Medline: 29104964]
45. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016 Jan;66:29-39 [FREE Full text] [doi: 10.1016/j.artmed.2015.09.007] [Medline: 26481140]
46. Hassanpour S, Langlotz CP. Unsupervised topic modeling in a large free text radiology report repository. J Digit Imaging 2016 Feb;29(1):59-62 [FREE Full text] [doi: 10.1007/s10278-015-9823-3] [Medline: 26353748]
47. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. AJR Am J Roentgenol 2017 Apr;208(4):750-753. [doi: 10.2214/AJR.16.16128] [Medline: 28140627]
48. Hassanpour S, Bay G, Langlotz CP. Characterization of change and significance for clinical findings in radiology reports through natural language processing. J Digit Imaging 2017 Jun;30(3):314-322 [FREE Full text] [doi: 10.1007/s10278-016-9931-8] [Medline: 28050714]
49. He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artif Intell Med 2019 Jan;93:43-49. [doi: 10.1016/j.artmed.2018.05.001] [Medline: 29778673]
50. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017;12(4):e0174708 [FREE Full text] [doi: 10.1371/journal.pone.0174708] [Medline: 28384212]
51. Jagannatha A, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proc Conf 2016 Jun;2016:473-482 [FREE Full text] [doi: 10.18653/v1/n16-1056] [Medline: 27885364]
52. Jonnagaddala J, Liaw S, Ray P, Kumar M, Dai H, Hsu C. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. Biomed Res Int 2015;2015:636371 [FREE Full text] [doi: 10.1155/2015/636371] [Medline: 26380290]
53. Karystianis G, Nevado AJ, Kim C, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res 2018 Mar;27(1) [FREE Full text] [doi: 10.1002/mpr.1602] [Medline: 29271009]
54. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning: a step towards automating medical concept extraction. J Am Med Inform Assoc 2016 Mar;23(2):289-296. [doi: 10.1093/jamia/ocv069] [Medline: 26253132]
55. Kim Y, Garvin J, Goldstein M, Meystre S. Classification of contextual use of left ventricular ejection fraction assessments. Stud Health Technol Inform 2015;216:599-603 [FREE Full text] [doi: 10.3233/978-1-61499-564-7-599] [Medline: 26262121]
56. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform 2015 Jun;55:188-195 [FREE Full text] [doi: 10.1016/j.jbi.2015.04.008] [Medline: 25917057]
57. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]
58. Lee H, Wu Y, Zhang Y, Xu J, Xu H, Roberts K. A hybrid approach to automatic de-identification of psychiatric notes. J Biomed Inform 2017 Nov;75S:S19-S27 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.006] [Medline: 28602904]
59. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 6;15:37 [FREE Full text] [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]
60. Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. J Am Med Inform Assoc 2016 Mar;23(2):387-395 [FREE Full text] [doi: 10.1093/jamia/ocv113] [Medline: 26521301]
61. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. Appl Clin Inform 2016 Jul 20;7(3):693-706 [FREE Full text] [doi: 10.4338/ACI-2016-01-RA-0015] [Medline: 27452794]

62. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58(Suppl):S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]
63. Liu S, Liu H, Chaudhary V, Li D. An infinite mixture model for coreference resolution in clinical notes. AMIA Jt Summits Transl Sci Proc 2016;2016:428-437 [FREE Full text] [Medline: 27595047]
64. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
65. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 2017 Nov;75S:S34-S42 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.023] [Medline: 28579533]
66. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015 Sep;22(5):1009-1019 [FREE Full text] [doi: 10.1093/jamia/ocv016] [Medline: 25862765]
67. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 2018 Jan 1;25(1):93-98 [FREE Full text] [doi: 10.1093/jamia/ocx090] [Medline: 29025149]
68. Maguen S, Madden E, Patterson OV, DuVall SL, Goldstein LA, Burkman K, et al. Measuring use of evidence based psychotherapy for posttraumatic stress disorder in a large national healthcare system. Adm Policy Ment Health 2018 Jul;45(4):519-529. [doi: 10.1007/s10488-018-0850-5] [Medline: 29450781]
69. Mai M, Krauthammer M. Controlling testing volume for respiratory viruses using machine learning and text mining. AMIA Annu Symp Proc 2016;2016:1910-1919 [FREE Full text] [Medline: 28269950]
70. Maldonado R, Goodwin T, Skinner M, Harabagiu S. Deep learning meets biomedical ontologies: knowledge embeddings for epilepsy. AMIA Annu Symp Proc 2017;2017:1233-1242 [FREE Full text] [Medline: 29854192]
71. Martinez D, Ananda-Rajah MR, Suominen H, Slavin MA, Thursky KA, Cavedon L. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. J Biomed Inform 2015 Feb;53:251-260 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.009] [Medline: 25460203]
72. Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc 2017 Apr 1;24(e1):e40-e46. [doi: 10.1093/jamia/ocw097] [Medline: 27413122]
73. Meystre S, Gouripeddi R, Tieder J, Simmons J, Srivastava R, Shah S. Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: a PHIS+ pilot study. J Med Internet Res 2017 May 15;19(5):e162 [FREE Full text] [doi: 10.2196/jmir.6887] [Medline: 28506958]
74. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. JMIR Public Health Surveill 2018 Apr 25;4(2):e29 [FREE Full text] [doi: 10.2196/publichealth.9361] [Medline: 29695376]
75. Napolitano G, Marshall A, Hamilton P, Gavin AT. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med 2016 Jun;70:77-83. [doi: 10.1016/j.artmed.2016.06.001] [Medline: 27431038]
76. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 2016 Nov;23(6):1077-1084 [FREE Full text] [doi: 10.1093/jamia/ocw006] [Medline: 27026618]
77. Patterson O, Forbush T, Saini S, Moser SE, du Vall SL. Classifying the indication for colonoscopy procedures: a comparison of NLP approaches in a diverse national healthcare system. Stud Health Technol Inform 2015;216:614-618. [Medline: 26262124]
78. Posada JD, Barda AJ, Shi L, Xue D, Ruiz V, Kuan P, et al. Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. J Biomed Inform 2017 Nov;75S:S94-104 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.019] [Medline: 28571784]
79. Rastegar-Mojarad M, Sohn S, Wang L, Shen F, Bleeker TC, Cliby WA, et al. Need of informatics in designing interoperable clinical registries. Int J Med Inform 2017 Dec;108:78-84 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.10.004] [Medline: 29132635]
80. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015 Dec;58(Suppl):S111-S119 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.010] [Medline: 26122527]
81. Roysden N, Wright A. Predicting health care utilization after behavioral health referral using natural language processing and machine learning. AMIA Annu Symp Proc 2015;2015:2063-2072 [FREE Full text] [Medline: 26958306]
82. Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med 2018 Mar 1;94:1-10 [FREE Full text] [doi: 10.1016/j.compbiomed.2017.12.026] [Medline: 29353160]

83. Scheurwegs E, Sushil M, Tulkens S, Daelemans W, Luyckx K. Counting trees in random forests: predicting symptom severity in psychiatric intake reports. J Biomed Inform 2017 Nov;75S:S112-S119 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.007] [Medline: 28602906]
84. Schuler A, Liu V, Wan J, Callahan A, Udell M, Stark D, et al. Discovering patient phenotypes using generalized low rank models. Pac Symp Biocomput 2016;21:144-155 [FREE Full text] [doi: 10.1142/9789814749411_0014] [Medline: 26776181]
85. Sevenster M, Buurman J, Liu P, Peters J, Chang P. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2015;6(3):600-610 [FREE Full text] [doi: 10.4338/ACI-2014-11-RA-0110] [Medline: 26448801]
86. Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. J Surg Res 2017 Mar;209:168-173 [FREE Full text] [doi: 10.1016/j.jss.2016.09.058] [Medline: 28032554]
87. Tahmasebi AM, Zhu H, Mankovich G, Prinsen P, Klassen P, Pilato S, et al. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. J Digit Imaging 2019 Feb;32(1):6-18 [FREE Full text] [doi: 10.1007/s10278-018-0116-5] [Medline: 30076490]
88. Tang B, Chen Q, Wang X, Wu Y, Zhang Y, Jiang M, et al. Recognizing disjoint clinical concepts in clinical text using machine learning-based methods. AMIA Annu Symp Proc 2015;2015:1184-1193 [FREE Full text] [doi: 10.1186/1472-6947-13-S1-S1] [Medline: 26958258]
89. Teixeira PL, Wei W, Cronin RM, Mo H, van Houten JP, Carroll RJ, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc 2017 Jan;24(1):162-171 [FREE Full text] [doi: 10.1093/jamia/ocw071] [Medline: 27497800]
90. Temple MW, Lehmann C, Fabbri D. Natural language processing for cohort discovery in a discharge prediction model for the neonatal ICU. Appl Clin Inform 2016;7(1):101-115 [FREE Full text] [doi: 10.4338/ACI-2015-09-RA-0114] [Medline: 27081410]
91. Torii M, Fan J, Yang W, Lee T, Wiley MT, Zisook DS, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform 2015 Dec;58(Suppl):S164-S170 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.011] [Medline: 26279500]
92. Tran T, Kavuluru R. Predicting mental conditions based on 'history of present illness' in psychiatric notes with deep neural networks. J Biomed Inform 2017 Nov;75S:S138-S148 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.010] [Medline: 28606869]
93. Trivedi G, Pham P, Chapman W, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc 2018 Jan 1;25(1):81-87 [FREE Full text] [doi: 10.1093/jamia/ocx070] [Medline: 29016825]
94. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging 2018 Apr;31(2):245-251 [FREE Full text] [doi: 10.1007/s10278-017-0021-3] [Medline: 28924815]
95. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]
96. Unanue IJ, Borzeshi EZ, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform 2017 Dec;76:102-109 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.007] [Medline: 29146561]
97. Walsh JA, Shao Y, Leng J, He T, Teng C, Redd D, et al. Identifying axial spondyloarthritis in electronic medical records of US veterans. Arthritis Care Res (Hoboken) 2017 Sep;69(9):1414-1420 [FREE Full text] [doi: 10.1002/acr.23140] [Medline: 27813310]
98. Wang C, Akella R. A hybrid approach to extracting disorder mentions from clinical notes. AMIA Jt Summits Transl Sci Proc 2015;2015:183-187 [FREE Full text] [Medline: 26306265]
99. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. J Am Med Inform Assoc 2015 Nov;22(6):1196-1204 [FREE Full text] [doi: 10.1093/jamia/ocv102] [Medline: 26232442]
100. Wang Y, Zheng K, Xu H, Mei Q. Clinical word sense disambiguation with interactive search and classification. AMIA Annu Symp Proc 2016;2016:2062-2071 [FREE Full text] [Medline: 28269966]
101. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 Jun 12;17(1):84 [FREE Full text] [doi: 10.1186/s12911-017-0483-8] [Medline: 28606174]
102. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. Crit Care Med 2018 Jul;46(7):1125-1132 [FREE Full text] [doi: 10.1097/CCM.0000000000003148] [Medline: 29629986]
103. Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 1;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]

104. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. AMIA Annu Symp Proc 2017;2017:1812-1819 [FREE Full text] [Medline: 29854252]
105. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Wang L, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 1;24(e1):e79-e86. [doi: 10.1093/jamia/ocw109] [Medline: 27539197]
106. Yadav K, Sarioglu E, Choi H, Cartwright WB, Hinds PS, Chamberlain JM. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. Acad Emerg Med 2016 Feb;23(2):171-178 [FREE Full text] [doi: 10.1111/acem.12859] [Medline: 26766600]
107. Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform 2015 Dec;58(Suppl):S171-S182 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.006] [Medline: 26375492]
108. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015 Dec;58(Suppl):S30-S38 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.015] [Medline: 26231070]
109. Ye Y, Wagner MM, Cooper GF, Ferraro JP, Su H, Gesteland PH, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PLoS One 2017;12(4):e0174970 [FREE Full text] [doi: 10.1371/journal.pone.0174970] [Medline: 28380048]
110. Yim W, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform 2016 Dec;64:179-191 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.005] [Medline: 27729234]
111. Yim W, Denman T, Kwan S, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. AMIA Jt Summits Transl Sci Proc 2016;2016:455-464. [doi: 10.1148/radiographics.21.1.g01ja18237] [Medline: 27570686]
112. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018 May;287(2):570-580. [doi: 10.1148/radiol.2018171093] [Medline: 29381109]
113. Zhang R, Ma S, Shanahan L, Munroe J, Horn S, Speedie S. Discovering and identifying New York heart association classification from electronic health records. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):48 [FREE Full text] [doi: 10.1186/s12911-018-0625-7] [Medline: 30066653]
114. Zhang E, Thurier Q, Boyle L. Improving clinical named-entity recognition with transfer learning. Stud Health Technol Inform 2018;252:182-187. [doi: 10.1007/978-3-319-96893-3_20] [Medline: 30040703]
115. Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Med Inform 2017 May 9;5(2):e12 [FREE Full text] [doi: 10.2196/medinform.7235] [Medline: 28487265]
116. Zheng J, Yu H. Assessing the readability of medical documents: a ranking approach. JMIR Med Inform 2018 Mar 23;6(1):e17 [FREE Full text] [doi: 10.2196/medinform.8611] [Medline: 29572199]
117. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying patients with depression using free-text clinical documents. Stud Health Technol Inform 2015;216:629-633. [doi: 10.3233/978-1-61499-564-7-629] [Medline: 26262127]
118. Knake LA, Ahuja M, McDonald EL, Ryckman KK, Weathers N, Burstain T, et al. Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data. BMC Pediatr 2016 Apr 29;16:59 [FREE Full text] [doi: 10.1186/s12887-016-0592-z] [Medline: 27130217]
119. Sohn S, Wang Y, Wi C, Krusemark E, Ryu E, Ali M, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc 2018 Mar 1;25(3):353-359. [doi: 10.1093/jamia/ocx138] [Medline: 29202185]
120. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
121. Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA--US Department of Veterans Affairs national-scale HIS. Int J Med Inform 2003 Mar;69(2-3):135-156. [doi: 10.1016/s1386-5056(02)00131-4] [Medline: 12810119]
122. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the veterans health administration. Health Aff (Millwood) 2014 Jul;33(7):1203-1211. [doi: 10.1377/hlthaff.2014.0054] [Medline: 25006147]
123. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
124. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]
125. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
126. Suominen H, Salanterä S, Velupillai S, Chapman W, Savova G, Elhadad N. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European

Languages. 2013 Presented at: CLEF'13; September 23-36, 2013; Valencia, Spain p. 212-231. [doi: 10.1007/978-3-642-40802-1_24]
127. Stubbs A, Uzuner O. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. J Biomed Inform 2015 Dec;58(Suppl):S20-S29 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.020] [Medline: 26319540]
128. Stubbs A, Uzuner O. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform 2015 Dec;58(Suppl):S78-S91 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.009] [Medline: 26004790]
129. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation. 2015 Presented at: SemEval'15; June, 2015; Denver, Colorado p. 806-814. [doi: 10.18653/v1/s15-2136]
130. Filannino M, Stubbs A, Uzuner O. Symptom severity prediction from neuropsychiatric clinical records: overview of 2016 CEGS N-GRID shared tasks Track 2. J Biomed Inform 2017 Nov;75S:S62-S70 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.017] [Medline: 28455151]
131. Stubbs A, Filannino M, Uzuner O. De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-18 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.011] [Medline: 28614702]
132. Sorace J, Aberle DR, Elimam D, Lawvere S, Tawfik O, Wallace WD. Integrating pathology and radiology disciplines: an emerging opportunity? BMC Med 2012 Sep 5;10:100 [FREE Full text] [doi: 10.1186/1741-7015-10-100] [Medline: 22950414]
133. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 2008 Mar;5(3):210-217. [doi: 10.1016/j.jacr.2007.09.004] [Medline: 18312970]
134. Ko YA, Hayek S, Sandesara P, Tahhan AS, Quyyumi A. Cohort profile: the emory cardiovascular biobank (EmCAB). BMJ Open 2017 Dec 29;7(12):e018753 [FREE Full text] [doi: 10.1136/bmjopen-2017-018753] [Medline: 29288185]
135. O'Leary KJ, Liebovitz DM, Feinglass J, Liss DT, Evans DB, Kulkarni N, et al. Creating a better discharge summary: improvement in quality and timeliness using an electronic discharge summary. J Hosp Med 2009 Apr;4(4):219-225. [doi: 10.1002/jhm.425] [Medline: 19267397]
136. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems collaboratory. J Am Med Inform Assoc 2013 Dec;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]
137. Ni Y, Kennebeck S, Dexheimer J, McAneney C, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
138. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 1;25(3):230-238. [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
139. Ho D, Liang E, Chen X, Stoica I, Abbeel P. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. In: Proceedings of the International Conference on Machine Learning. 2019 Presented at: CML'19; June 10-15, 2019; Long Beach, USA p. 2731-2741 URL: http://proceedings.mlr.press/v97/ho19b/ho19b.pdf
140. Li Y, Cohn T, Baldwin T. Robust Training under Linguistic Adversity. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017 Presented at: ACL'17; April 3-7, 2017; Valencia, Spain p. 21-27 URL: https://www.aclweb.org/anthology/E17-2004/ [doi: 10.18653/v1/e17-2004]
141. Xie Z, Wang SI, Li J, Levy D, Nie A, Jurafsky D, et al. Data Noising as Smoothing in Neural Network Language Models. In: Proceedings of the 5th International Conference on Learning Representations. 2017 Presented at: CLR'17; April 24-26, 2017; Toulon, France URL: https://nlp.stanford.edu/pubs/xie2017noising.pdf
142. Kobayashi S. Contextual Augmentation: Data Augmentation by Words With Paradigmatic Relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: ACL'18; 2018; New Orleans, Louisiana, USA p. 452-457 URL: https://www.aclweb.org/anthology/N18-2072/ [doi: 10.18653/v1/n18-2072]
143. Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-IJCNLP; 2019; Hong Kong, China p. 6382-6388 URL: https://www.aclweb.org/anthology/D19-1670/ [doi: 10.18653/v1/d19-1670]
144. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010 Oct;22(10):1345-1359 [FREE Full text] [doi: 10.1109/tkde.2009.191]
145. Do C, Ng A. Transfer learning for text classification. Adv Neural Inf Process Syst 2006;18:299-306 [FREE Full text]
146. Mintz M, Bills S, Snow R, Jurafsky D. Distant Supervision for Relation Extraction Without Labeled Data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2. 2009 Presented at: ACL'09; 2009; Suntec, Singapore p. 1003-1011 URL: https://www.aclweb.org/anthology/P09-1113/ [doi: 10.3115/1690219.1690287]

Abbreviations

CDW: Corporate Data Warehouse
EHR: electronic health record
ICD: International Classification of Diseases
IE: information extraction
MeSH: Medical Subject Headings
MIMIC: Medical Information Mart for Intensive Care
MRI: magnetic resonance imaging
NER: named entity recognition
NLP: natural language processing
RQs: research question
VHA: Veterans Health Administration
VistA: Veterans Information Systems Technology Architecture
WSD: word sense disambiguation

Edited by M Focsa, G Eysenbach; submitted 28.01.20; peer-reviewed by R Stewart, K Chen, C Lovis; comments to author 21.02.20; revised version received 24.02.20; accepted 24.02.20; published 31.03.20

Please cite as:
Spasic I, Nenadic G
Clinical Text Data in Machine Learning: Systematic Review
JMIR Med Inform 2020;8(3):e17984
URL: http://medinform.jmir.org/2020/3/e17984/
doi: 10.2196/17984
PMID: